ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse50922020.pdf
https://doi.org/10.30534/ijatcse/2020/50922020

On Feature Selection Stability and Privacy Preserving Data Mining: A Data Perspective

MohanaChelvan P[1], Dr. Perumal K[2]
[1]Department of Computer Science, Karpagam Academy of Higher Education (KAHE), Coimbatore, India, Email: pmohanselvan@rediff.com
[2]Department of Computer Applications, Madurai Kamaraj University, Madurai, India, Email: perumalmala@gmail.com

ABSTRACT

The data relating to data mining has turned out to be profoundly multi-dimensional in the recent past. It is also to be noted that such dimensionality has rapidly expanded over time. Moreover, in light of the positive assessment norms for enhanced data mining concerts, feature selection opts for a petite subset of the significant features from the original dataset. The stability of the feature selection is a key criterion in feature selection algorithms. Moreover, the most important aspect is its sturdiness in reducing the disturbances in the training data or in the expansion of the most recent samples. Lately, it has been demonstrated that the stability of the feature selection usually centers on data, and that it is not entirely unbiased in terms of algorithm. The privacy-preserving data mining changes a portion of the sensitive and quasi-identifying attributes in order to keep the conceivable re-identification of an individual's tuple through intrusive or malignant data miner and brings a choppy privacy conserved dataset. Since the stability of the feature selection relies primarily upon data, the stability of the feature selection lessens with such privacy-preserved choppy datasets. Besides, the privacy preserving ruffling associates stresses the stability of the selection of features and data utility. Picking proper privacy-preserving data mining technique with significant privacy-preserving ruffling to enhance feature selection stability alongside the greater privacy-preservation and data utility is consequently a challenging issue in the field of research. Hence, the present paper intends to highlight the issue with reference to the three algorithms for privacy-preserving data mining and their relative analysis.

Key words: High dimensional data, Feature Selection, Selection Stability, Stability Measures, Privacy Preservation, Kuncheva Index KI

1. INTRODUCTION

Data mining might be phenomenal by the analysis of chronicled organizational datasets. It can significantly extricate the already hard to perceive, non-immaterial, undeniable and spellbinding patterns or expertise. Data mining is crucial for organizations for accomplishing an edge over their contenders [1]. Because of the present signs of progress in information machinery that encompasses web predicated business and e-administration, there can be an explosion of microdata about citizenry which is typically high dimensional, and such a kind of issue is better known as the "curse of dimensionality" [2]. Therefore, the feature selection is employed as a unique dimension diminution approach for the existent high dimensional microdata [3, 4].

Dimensionality lessening techniques are predominantly categorized as extraction of features and the selection of features. In the approach of extracting features, the features are usually projected into a space with lower dimension. The feature selection approach, on the other hand, seeks to select a minimal subset of pertinent traits that minimize redundancy and maximize the association with the target i.e., the class label. While equaling these two techniques, the selection of features is superior in terms of greater readability and interpretability because it preserves the original feature values in a shortened space, whereas the extraction of features transmutes the data from the original space into a different space with a relatively lower dimension that cannot be concomitant to the traits of the original space. Consequently, further analysis of the new space thus becomes quite challenging and also problematic when considering the fact that the remodelled features obtained from the extraction technique also have no physical implications. Therefore, the feature selection technique has been implemented in the experimentations of the present paper.

Feature selection eliminates incongruous, unnecessary, and noisy data. Hence the reduced subset of relevant traits works considerably better than the complete set of high dimensional microdata. The benefits of selecting traits include a better investigation of precision in terms of predictive accuracy, diminished computational costs, better interpretability of the model, result from comprehensibility, and also decreased stockpiling space [3, 4]. Moreover, feature selection amends with privacy safeguarding of citizenry's records by improving data de-identification, jettisoning several quasi-identifier traits, and by feasibly confiscating extraneous and noisy traits.

1.1. Feature Selection Stability

Feature selection stability enables the feature selection algorithm to choose near or comparative subsets of traits in subsequent algorithmic cycles for the extension or erasure of some tuples from the dataset [5]. The stability of the feature selection is critical for data mining since the miscellaneous diverse subsets of significant traits in each new experimental release ultimately prompts an uncertain conclusion towards experimental consequences [6-8]. The researchers especially pay attention to the significance of the stability in the selection of features since it has attenuated the previously established convictions in drawing conclusions. The results of unstable feature selection, also to a considerable extent, have an impact upon the researchers wherein it both tapers the self-confidence and let them remain indecisive in drawing there search findings. Furthermore, feature selection stability is currently acknowledged as a key standard for feature selection algorithms, thus making it a hotly debated issue in research [7, 9].

1.2. Data Perspective nature of Feature Selection Stability

So far feature selection stability was contemplated from an algorithmic perspective in previous researches; however, it overlooked the ultimate qualities of the dataset, which significantly affect the stability of the selection of features. Likewise, more recently, a considerable attention has been given to the data perception and the stability of the selection of features. The stability of feature selection algorithms generally relies upon the sensitivity of feature selection and data discrepancies [7, 10]. It is moreover found [11] that the underlying physiognomies of the data can significantly influence the stability of the feature selection algorithms and that the issue of stability can to a great extent be data-reliant, at the same time, not entirely algorithmic impartial. Several researchers as of late have realized that stability of the selection of features hinges on for the most part on data, i.e., dataset characteristics and data variance [11-16]. Besides, the data discrepancy is typically instigated by noise. The components that sway stability of the selection of features incorporate the number of chosen traits [7], such as, dimensionality, sample size [6] and an assortment of data dissemination crosswise over various folds.

2. RESEARCH PROBLEM

Unstable feature selection outcomes prompt inconsistent conclusions and consequently lessen the self-assurance of researchers. In addition, it is to be noted that some outstanding features of selection algorithms is that it identifies all the ill effects and issues of low stability so as to also ensure the instauration of minimal data disruption in training set. The issues of feature selection stability are enhancing, particularly for choppy datasets where the composition of the original dataset is disturbed to a certain degree, for example, in incremental datasets, privacy preserved datasets, as stability is by and large data-centric. Changing of datasets can be contemplated instantly by erasing or including tuples, adding noise to the traits at feature level or by consolidating both. Furthermore, privacy preservation safeguards the privacy of an individual by amending sensitive and quasi-identifying traits of the dataset. Also, privacy preserving alteration ought not to enormously influence the feature selection stability, since the stability mostly depends upon the data.

In the case of choppy privacy preserved datasets, data discrepancy is identified to be the foremost source that leads to the instability of feature selection. Moreover, the contrary correlation helps attain the techniques for accomplishing great stability in the selection of features and safeguard privacy. The stability of the feature selection for the choppy privacy preserved datasets is crucial for better data utility, as accuracy is emphatically linked with the stability of the feature selection. The intensified feature selection stability issues of the choppy privacy preserved datasets, for better data utility, and safeguarding the privacy of an individual's records are the challenging research problems which is to be analysed in the present paper.

3. METHODOLOGY

There are a variety of privacy preserving data mining techniques and a series of ways to protect the privacy of individuals in data mining, which includes deletion, swapping, aggregation, noise addition, and encryption. Data publishing methods such as k-anonymity, l-diversity, t-closeness, slicing and differential privacy use techniques such as bucketization and generalization. The diverse privacy preserving data mining techniques are assessed for their strengths and defects. Based on the assessment, new privacy preserving data mining algorithms are developed. The developed privacy preserving data mining algorithms are implemented in experimental datasets, so as to alter the sensory attributes of the datasets and to protect the privacy of the citizenry. Similarly, the feature selection stability of the algorithms is evenly disturbed since it usually depends upon the datasets. The privacy preserving data mining algorithms should, therefore, preserve the privacy of the citizenry, and moreover, should have greater feature selection stability as well as accuracy.

Three new privacy preserving data mining algorithms are created by using several privacy preservation techniques. The proposed methodology for privacy preserving data mining algorithms is shown in Fig. 1.

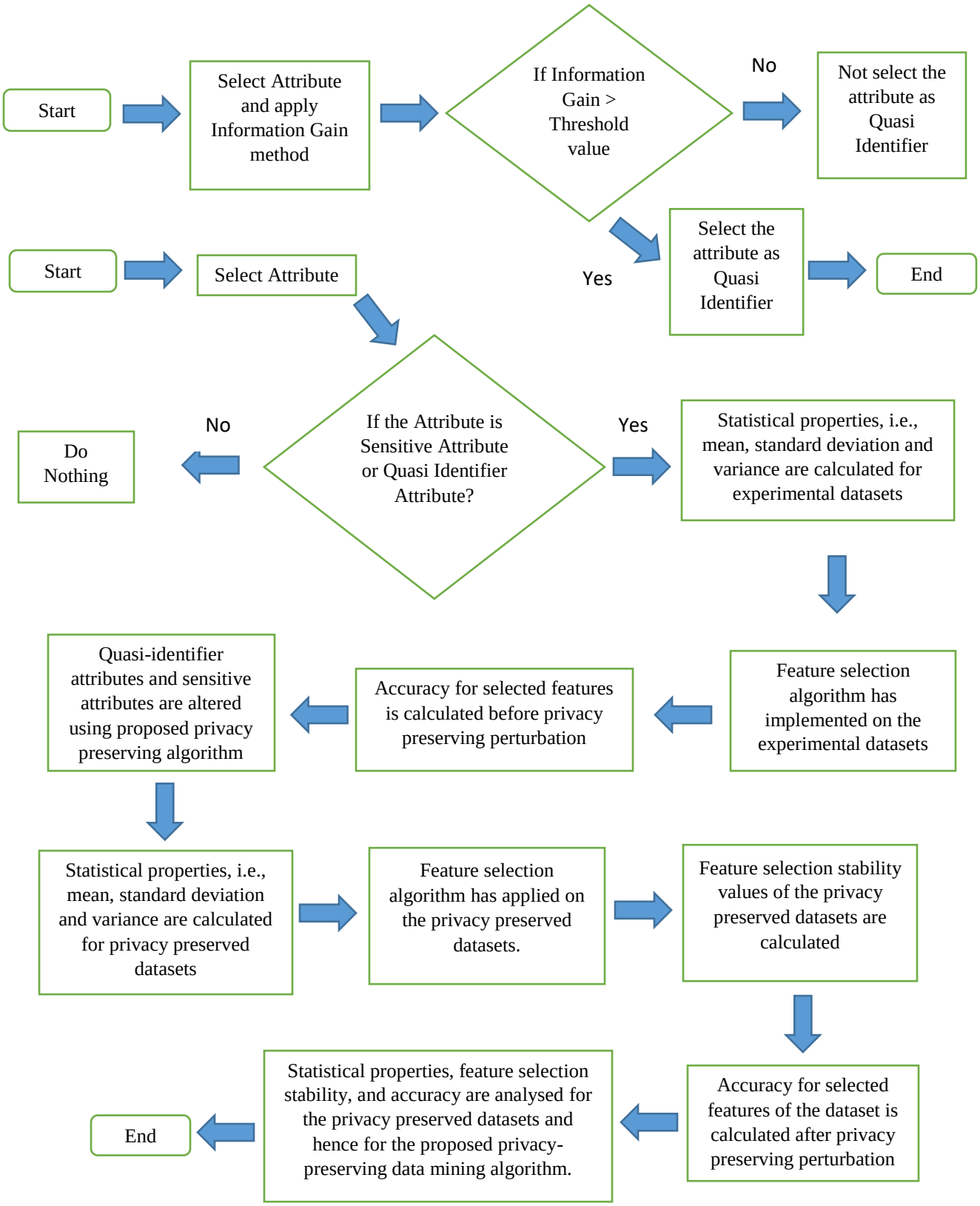


Figure 1: Block diagram of the proposed methodology

4. FEATURE SELECTION ALGORITHMS

The feature selection technique, as a rule, underpins the following three methodologies: filter, wrapper, and hybrid [17]. The filter approach of feature selection is employed to obliterate features in specific measures or criteria and subsequently, the goodness of a feature is assessed by utilizing the inherent or statistical properties of the dataset. A feature is chosen based on the most appropriate feature bolsters, and its properties for both machine learning and data mining applications. Inside the wrapper approach, a set of features is created and the goodness of the set is assessed utilizing a specific classifier. In due course, the methodology keeps intact that the point of some classifier is to classify the features inside the dataset, and to bolster this rank, a feature is picked from the predetermined application. The embedded model incorporates the benefits of every one of the above models. The hybrid approach exploits the two methodologies by taking advantage of the different analytical criteria in various search stages. The feature selection algorithms employed in the investigations are Information Gain IG [18] and Correlation-based Feature Selection CFS [6].

4.1. Information Gain IG

Entropy is a criterion to analyse and test the impurity ofanactual training set. It is depicted asa symmetric measure of the information Y given by X, which symbolizes the number by the decline of Y [18]. The measure is comprehended as IG given in (1).

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \quad (1)$$

IG could be symmetrical measure because the information increased in X once it is perceptive of Y juxtaposes with the increase of Y with the increase of X.IG has a drawback that it favours features with extra values regardless of whether they appear to be extra informative. The information picked up by reference to the class is estimated by the assessment of the value of an attribute. The independence between a feature and the class label is estimated by IG, and considers the distinction between the entropy of the feature and the conditional entropy indicated by the class label as in (2).

$$IG(Class,Attribute) = H(Class) - H(Class|Attribute) \quad (2)$$

4.2. Correlation-based Feature Selection CFS

The value of a set of attributes is evaluated by CFS by considering the degree of repetition among them, notwithstanding the individual prognostic ability of each feature. Subsets of features that have low interrelationships among categories, however that are incredibly related among the classes, are most favoured [6]. CFS explains the least complex feature set and can be joined with search techniques to admire bi-directional search, backward elimination, forward selection, genetic search, and bestfirst search. Authors ought to have GA as a search approach with CFS as fitness.

$$r_{zc} = \frac{k\, r_{zi}}{\sqrt{k + (k-1)\, r_{ii}}} \quad (3)$$

For instance, r_{zc} is that the relationship between the class variable and the summed subset features, k is that of the subset features, r_{zi} is that the average of the correlations between the class variable and the subset features, and r_{ii} is that of the average inter-correlation between the subset features [8].

5. SELECTION STABILITY MEASURES

Stability may only be estimated by the comparison between the result outcomes. Hence, stability is more important if the similitude is more prominent. Stability by index, stability by rank, and stability by weight are the three principal categories of stability guesstimates that rely upon the portrayal of the selection technique yield [7]. In stability by index, the indices of the chose features are taken into consideration. In the aforementioned class, the chose features does not have an unequivocal order or relating relevant weight. In the class of stability by rank, the hierarchical list of features influences the stability analysis. In stability by weight class, each feature has apportioned a weight consistent with the associate degree.

Each of the three stability classes is produced by the weights of the respective features. In any given case, each domain focuses upon a totally different output at the same time encourages a particular stability for a precise output. It is also important to extend that a similar rank does not have consistent weight and equivalent assigned subset.

The three necessities deemed necessary for the estimation of stability [19] are as per the following:

- Monotonicity: If there is an enormous overlap between the chose subsets, the outcome should be of horrific stability.

- Limits: The consequence of every stability estimation technique ought to be limited among constants comparing to [-1, 1] or [0, 1]. These limits are autonomous of any dataset factor just as the dimensionality of the dataset m or the number of chose features k. These cut-off points ought to be at any rate

once the sets are flimsy and moved toward becoming the most indistinguishable or stable.

- Correction for chance: The measure should have a tireless rectification that prompts an intersection happening inadvertently because of the high-dimensional chose subset. The bigger the cardinality of the chose subsets, the more noteworthy the likelihood for a bigger intersection between the subsets are.

A number of stability measures have been offered to measure the comparability amongst the chosen feature subsets [7, 10, 19, and 20]. The Kuncheva Index KI [20] is an index-predicated measure of stability. The other stability measures in the class are the Jaccard Index [7], the Dice Index [10], and the Average Normal Hamming Distance AHND [19]. The Spearman's Rank Correlation Coefficient SRCC [7] is a member of the rank-predicated stability class. Pearson's Correlation Coefficient PCC [7] is the stability valuation of the weight classification.

5.1. Stability by index

The chose subset features in the stability by index class are portrayed either as a vector of indices reminiscent of the chose features or as a binary vector with cardinality equivalent to m, wherever fi = 1 means the i^{th} feature is chosen. In this class of estimation, it is conceivable to be dealt with various chose features among k and m, yet with the rank or weight estimation, it is beyond the realm of imagination. Consequently, these estimations have regular outcome confines wherever some are inside the interim [0, 1] or [-1, 1], though others are not limited in the least degree. In a nutshell, stability by index estimation class estimates the number of overlaps within which the stability is estimated.

5.1.1. Kuncheva Index KI

The two subsets of traits retain a lion's share in the stability measures. The noteworthy cardinality is that the list of chosen traits is fervently consistent with the threat of overlap. The Kuncheva Index KI is proposed in [20] to defeat the aforesaid detriment, and the latter incorporates an adjustment term to stay away from the intersection, which is unintentional among the chosen subsets of traits. KI is the only measure that agrees to every one of the necessities affirmed in [20], i.e., monotonicity, limitation, and chance correction.

$$KI\,(F'_1, F'_2) = \frac{\left|F'_1 \cap F'_2\right| . m - k^2}{k\,(m - k)} \qquad (4)$$

The KI results extend [-1, 1], wherever - 1 implies that there is no intersection between the lists and k= m/2. KI achieves 1 when Ƒ'1 and Ƒ'2 are indistinguishable, which implies that the cardinality of the intersection set is k. KI esteems are fitting for severe drowning lists at the point of zero. In the Jaccard Index and most of the other measures, the number of features that has chosen k will have an impact on stability. It gives higher stability esteems once k gets bigger and closer to m. The revamping of the term is structured in KI so that it eventually finishes with an appeal. Despite that, KI does not experience the ill effects as a result of the correction term, which gives negative weight to k. Hence, the Kuncheva Index was subsequently applied in the experimentations.

6. PRIVACY PRESERVING DATA MINING

By all means, the standard conjecture in privacy preserving data mining is that, the candidate data to be revealed comprises multiple forms of traits. Parenthetically, Burnett et al., [21] outline the ensuing types:

- Identifiers - attributes that unambiguously resolve an individual like name, social security number, toiler id etc.

- Quasi-identifiers - non-sensitive attributes like age, zip code, or gender that would be practiced in linkage attacks.

- Non-sensitive attributes are deficit of any common attributes and they are neither quasi-identifiers nor sensitive.

- Sensitive attributes - personal attributes are quite private in nature. For example, diseases or financial gain.

Privacy preservation in data mining ensures confidentiality and to safeguard sensitive data against undesirable revelation. Several countries put regulation on processing and usage of personal data. For example, The European Union has sanctioned a new Regulation that lays down rules involving to the safeguard of ordinary persons with regard to the processing of personal data and rules relating to the free movement of personal data [22]. There will be need for information security, analysis of human factors in information security and review of human errors in information security contexts [23].

The technique offers individual privacy and furthermore empowers data mining for extricating helpful knowledge from the available data. In data mining, touchy raw data and tactile knowledge of data mining outcomes are in some way or the other guaranteed by the contortion of the original dataset, by employing the created algorithms [24]. High data quality and privacy are the indispensable necessities of privacy conservation techniques. To ensure data sensitive to

privacy, the dataset must be disturbed in a couple of ways [3] because the stability of the selection of features depends upon on the data to a great extent.

A wide scope of privacy preservation strategies is arranged and predicated on basis of distortion, clustering, associative classification, outsourced data mining, association rule, hide association rule, taxonomy and are uniformly distributed [25, 26]. There are numerous approaches to bother data for privacy conserving data mining. They include: suppression, ruffling, noise addition, rounding or coarsening, data swapping, aggregation, data shuffling, microaggregation, and encryption. On the other hand, the data publishing strategies involves k-anonymity, l-diversity, t-closeness, slicing, and the differential privacy use methods comprise of bucketization and generalization. Every technique has its own merits and demerits that are imperative for the stability of the selection of features and the data utility.

In this paper, there have been three privacy conserving data mining algorithms, Algorithm 1, Algorithm 2 and Algorithm 3, are employed. Algorithm 1 depends on data distortion methods and Algorithm 2 depends on slicing procedures, though Algorithm 3 depends on dogged noise addition strategies.

6.1. Algorithm 1using Data Distortion Techniques

The planned privacy protection algorithm employed in the present investigation has been presented in Fig. 2. Pseudo code for Algorithm 1 using Data Distortion Techniques is shown in Fig. 3. The data distortion procedure, otherwise called data ruffling or data randomization technique incorporates, additive data ruffling, multiplicative data ruffling, rotational data ruffling, and the methodology of the reconstruction tree. The data distortion procedure adjusts the touchy attributes within which their original qualities cannot be distinguished. However, their global properties continue as before and remain unaltered. The privacy protection algorithms that are reliant on mutilation are generally data concealing procedures, yet they do not govern any strategies of concealment. The privacy conserving algorithm toils at the single trait level. The data distortion technique exploitsthe analysis of association rules, followed by classification and clustering. The privacy-preserving data mining algorithm uses data distortion techniques combined with data shuffling, micro aggregation, value swapping, noise addition, and rotational techniques.

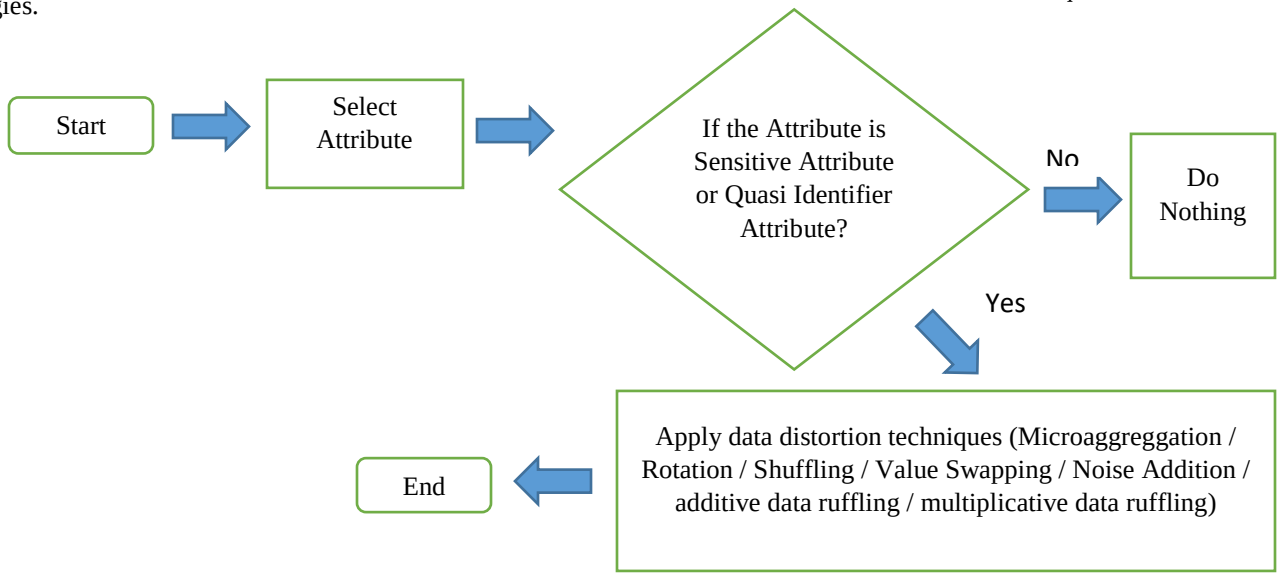


Figure 2: Proposed Methodology for Algorithm 1 using data distortion technique

Input: Dataset with Selected Sensitive Attribute and Quasi identifier Attribute values
Output: Apply Privacy Preserving Perturbation and return Dataset with perturbed Attribute value
Start
Select Attribute
If the selected attribute is not sensitive or quasi identifier attribute Do nothing
Else Apply data distortion techniques (Microaggreggation / Rotation / Shuffling / Value Swapping / Noise Addition) according to the attribute domain values
End

Figure 3: Pseudo code foe Algorithm 1 using Data Distortion Techniques

6.2. Algorithm 2using Slicing Technique

The planned privacy preservation algorithm operated in the trials is depicted in Fig. 4. Pseudo code for Algorithm 2 using Slicing Technique is shown in Fig. 5. The slicing technique was employed in a blend with value swap and suppression procedure amid the aforementioned algorithm. The slicing procedure coordinates superbly with high-dimensional microdata [27]. The slicing procedure breaks the microdata dataset vertically and on a level plane. Extremely correlated attributes are put in the slice block, and immaterial attributes are divided. The slicing strategy is more helpful than various procedures like bucketization and generalization. Besides that, the slicing procedure has an issue in uncovering certainties about the citizenry.

Slicing produces invalid records because of the breakdown of records that will bring about a diminution in the data utility. An association of invalid traits can deliver invalid records all through the slicing procedure, by bringing about the revelation of privacy-sensitive data. Similarly, upon slicing, if there is a meaningless connection between the tuple values, the corresponding values are translated into purposeful values for greater data utility. Negative association rules must be recognized to extricate invalid records. The strategy of injection was exploited in [28] to diagnose negative association rules. Value swapping was applied to modify the slicing method for a negative association and background attack, and together the data utility is moved forward. Suppression or deletion technique has been considerately implemented to prevent the disclosure of sensitive data. Since the slicing strategy cannot anticipate identity revelations and for the expanded privacy of sensitive tuples, the generalization of columns with suppression rehearses was carried out.

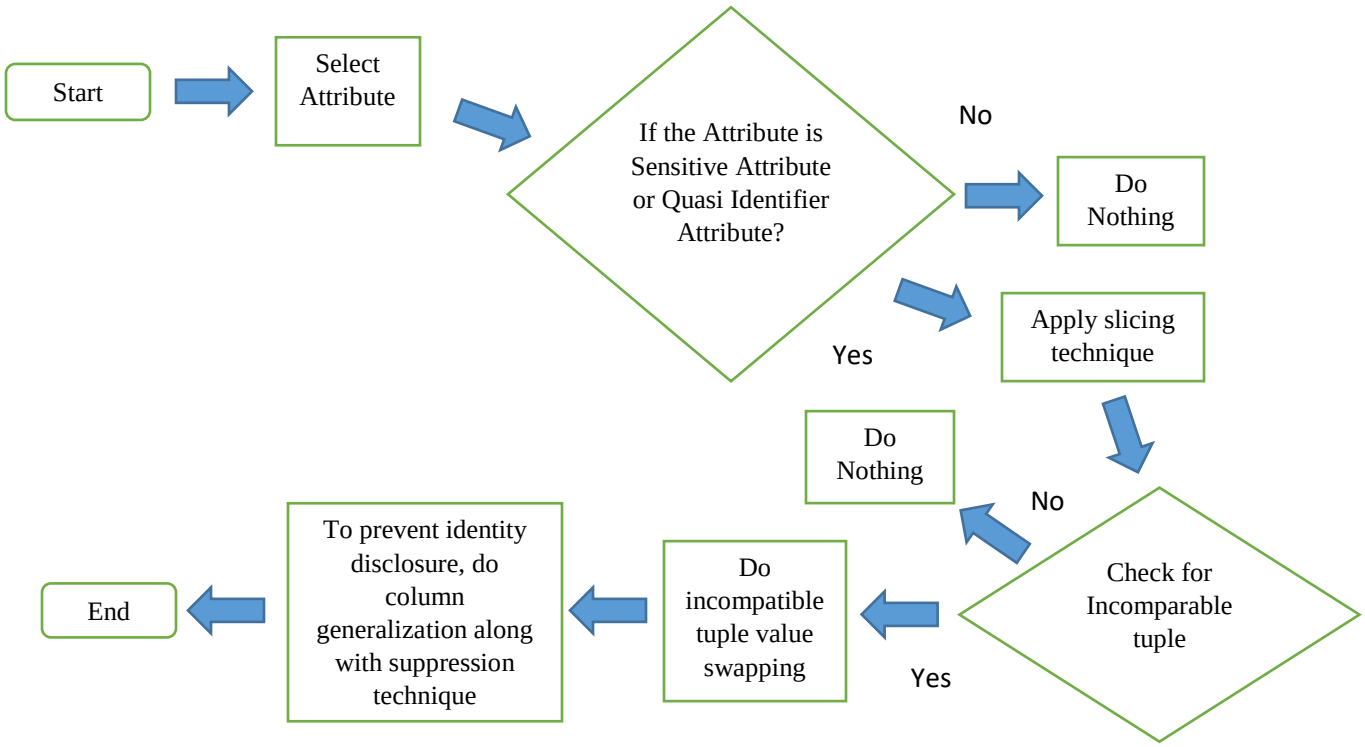


Figure 4: Proposed Methodology for Algorithm 2 using slicing technique

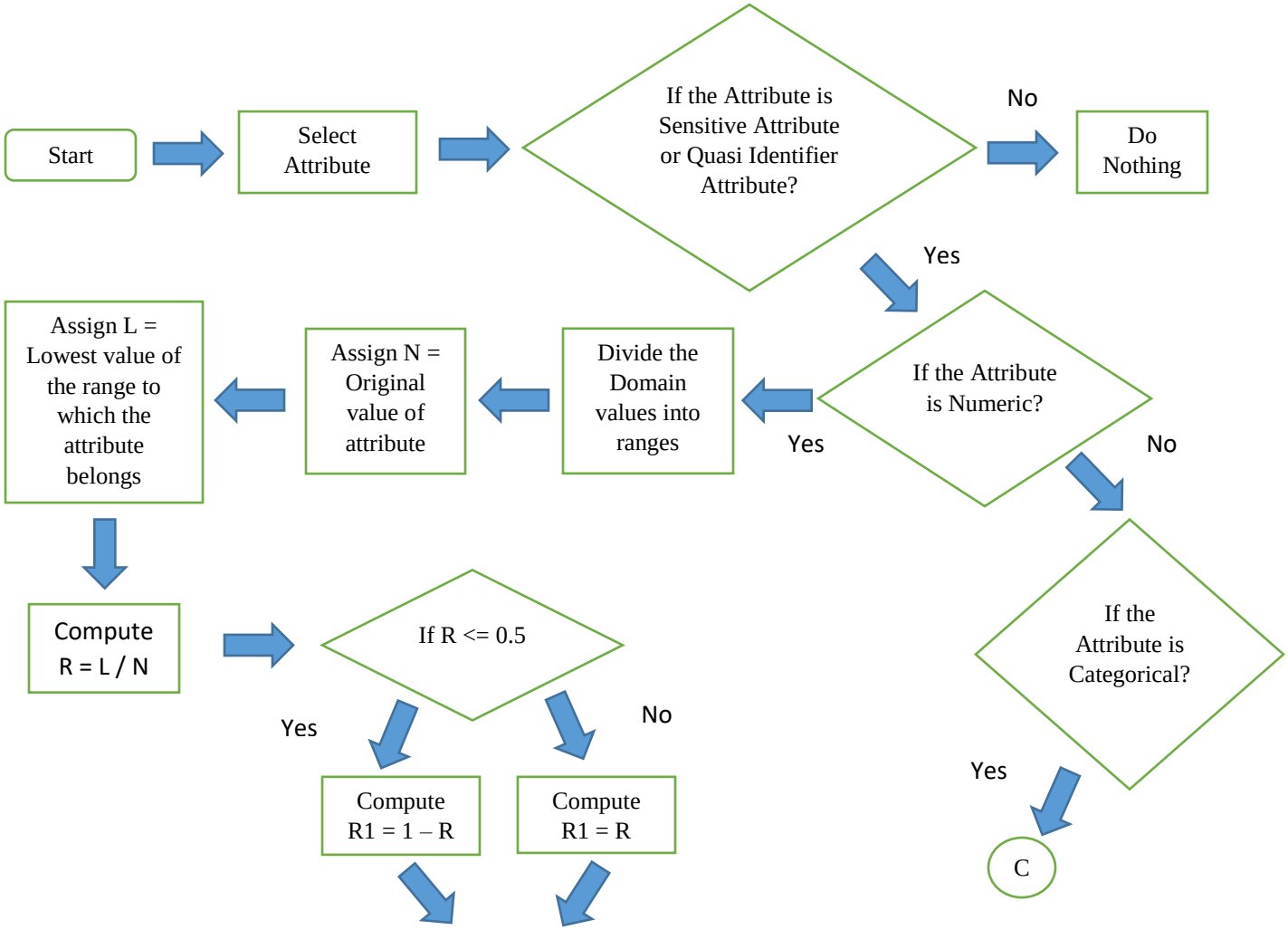
Input: Dataset with Selected Sensitive Attribute and Quasi identifier Attribute values
Output: Apply Privacy Preserving Perturbation and return Dataset with perturbed Attribute value
Start
Select Attribute

If the selected attribute is not sensitive or quasi identifier attribute Do nothing
Else Apply slicing technique
If the tuple is not Incompatible tuple D nothing
Else Do incompatible tuple value swapping
To prevent identity disclosure, do column generalization along with suppression technique
End

Figure 5: Pseudo code for Algorithm 2 using Slicing Technique

6.3. Algorithm 3 using Calculated Noise Addition Technique

The projected privacy protection algorithm [29] applied in the trials has been represented in Fig. 6. Pseudo code for Algorithm 3 using Calculated Noise Addition Technique is shown in Fig. 7. Amidst the proposed algorithm, the reckoned noise was inserted for the domain values of numerical traits, while the generalization methodology was kept functional for the domain values of categorical traits with respect to the known quasi-identifiers and delicate attributes. The algorithmic program intended for data mining for privacy preservation shields delicate data with influential feature selection stability and accuracy, and it can fabricate a model for more conspicuous data utility.



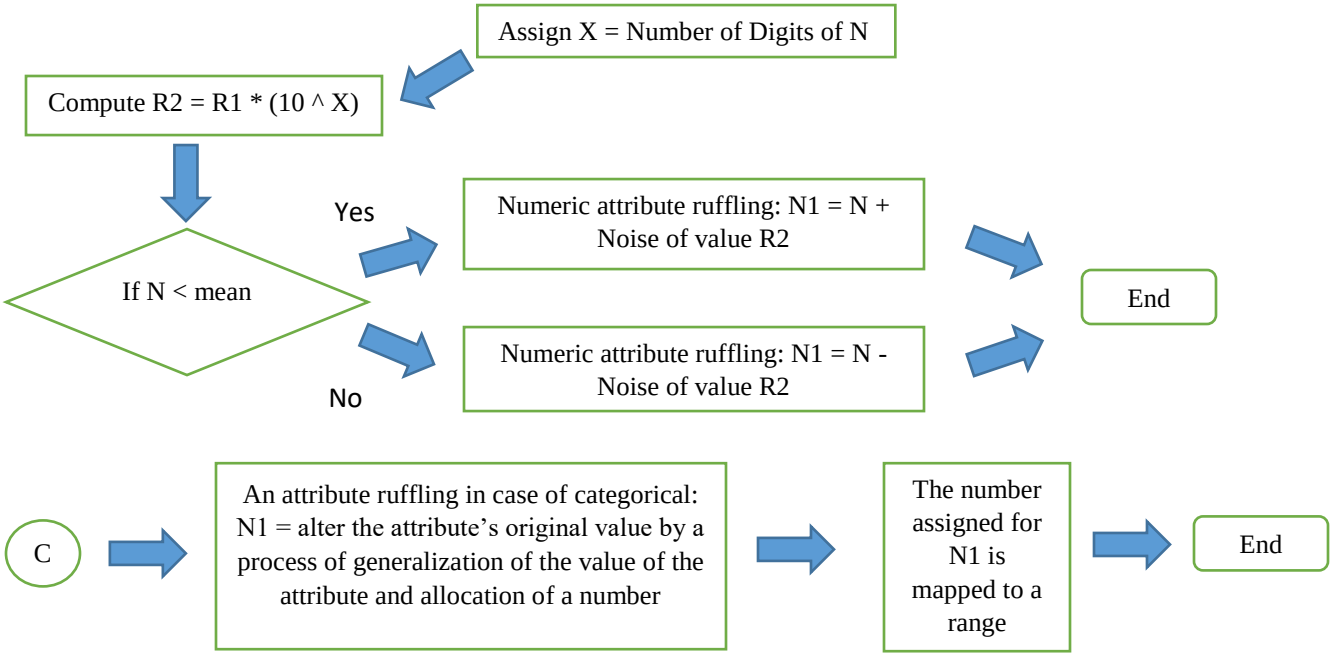


Figure 6: Methodology of proposed algorithm 3 using calculated noise addition technique

Input: Dataset with Selected Sensitive Attribute and Quasi identifier Attribute values
Output: Apply Privacy Preserving Perturbation and return Dataset with perturbed Attribute value
Start
Select Attribute
If selected attribute is not Sensitive attribute or Quasi-identifier attribute
Do Nothing
Else
If Attribute is Categorical Attribute go to Categorical Pseudo code
Else
If Attribute is Numeric
Divide the Domain values into Ranges
Assign Original value of Attribute to N
Assign lowest value of the range to which the attribute belongs to L
Divide L by N and assign the value to R
If R value is Greater than 0.5
Assign 1 – R value to R1
Else
Assign R value to R1
Assign number of digits of N to X
Compute R1 multiplied by 10 to the power of X and assign the value to R2
If N value is greater than mean of the selected range
Add noise value of R2 to N as N1 which is perturbed value of numeric attribute value
Else subtract noise value of R2 to N as N1 which is perturbed value of numeric attributevalue
End
Categorical Pseudo code

Start
Alter the attribute's original value by a process of generalization of the value of the attribute N
Allocation of a number for the range to N1
The number assigned for N1 is mapped to a range
End

Figure 7: Pseudo code for Algorithm 3 using Calculated Noise Addition Technique

7. EXPERIMENTS

Various data mining tools are available in the market including WEKA, R, Orange, Rapid Miner and Tanagra [30]. In the experiments, R is used as a data mining tool.

7.1. Methodology

The planned methodology for the privacy safeguarding algorithms is as per the following:

a) Quasi-identifiers are chosen to make use of the ranking strategy for information gain.
b) Statistical properties for trial datasets, for example, mean, standard deviation and variance are dogged.
c) For trial datasets, the CFS feature selection algorithm was linked.
d) The accuracy for chosen traits is dogged before privacy-protecting ruffling.
e) Quasi-identifier attributes and sensitive attributes are disrupted by the privacy conserving algorithm.
f) For privacy-conserved datasets, factual properties, for example, mean, standard deviation and variance are dogged.
g) The CFS feature selection algorithm has been linked to privacy preserved datasets.
h) Feature selection stability guesstimates of the privacy conserved datasets are dogged by employing Kuncheva Index KI.
i) The accuracy for chosen traits is dogged after privacy conserving ruffling.
j) Statistical properties, feature selection stability, and accuracy are analysed for the datasets conserved for privacy and for the privacy conservation algorithms.

7.2. Datasets Used

The two datasets employed in the investigations are the Census-Income (KDD) [31] and the Insurance Company Benchmark (COIL 2000) [31] datasets. The datasets are acquired from the KEEL dataset repository. Table 1 demonstrates the datasets virtues. Within the chronicled datasets, the census dataset has each categorical and numerical values, while the Coil 2000 dataset simply has numerical values.

Table 1: Characteristics of datasets Census and Coil 2000

S. No.	Datasets Characteristics	Datasets	
		Census	Coil 2000
1	Type	Classification	Classification
2	Origin	Real World	Real World
3	Instances	143228	9782
4	Features	41	85
5	Classes	3	2
6	Mislaid Values	Yes	No
7	Attribute Type	Numerical, Categorical	Numerical

7.3. Result Analysis

The graded traits are accomplished by reckoning the information gain with admiration for the class, and by taking into account the significance of a trait. It was consummated with the Information Gain IG feature selection algorithm. Upheld by the graded attributes, the quasi-identifiers are determined and selected for privacy conserving ruffling along with prickly attributes. The algorithms Algorithm 1, Algorithm 2, and Algorithm 3 perturb the quasi-identifiers and sensitive traits of the trial datasets. Each domain value of the chosen trait has been changed for a hundred percent privacy preservation so that a trespasser or mischievous data miner even with noteworthy background information cannot be positive about the precision of a re-identification.

The Fig. 8, underneath exhibits the statistical properties of the attribute 'Atr 0' of the original and perturbed

experimental dataset Census which has been developed. From the chart, it has been known that there are minor alterations in the measures of the statistical properties of the original and perturbed data sets for the attribute 'Atr 0' due to privacy preserving perturbation.

For instance, Fig. 9, presented underneath exhibits the histogram of the attribute ' Atr 0' of the original and modified test dataset Census that operates Algorithm 3. The diagram demonstrates that the recurrence of the trait values of the original and changed datasets as almost equivalent.

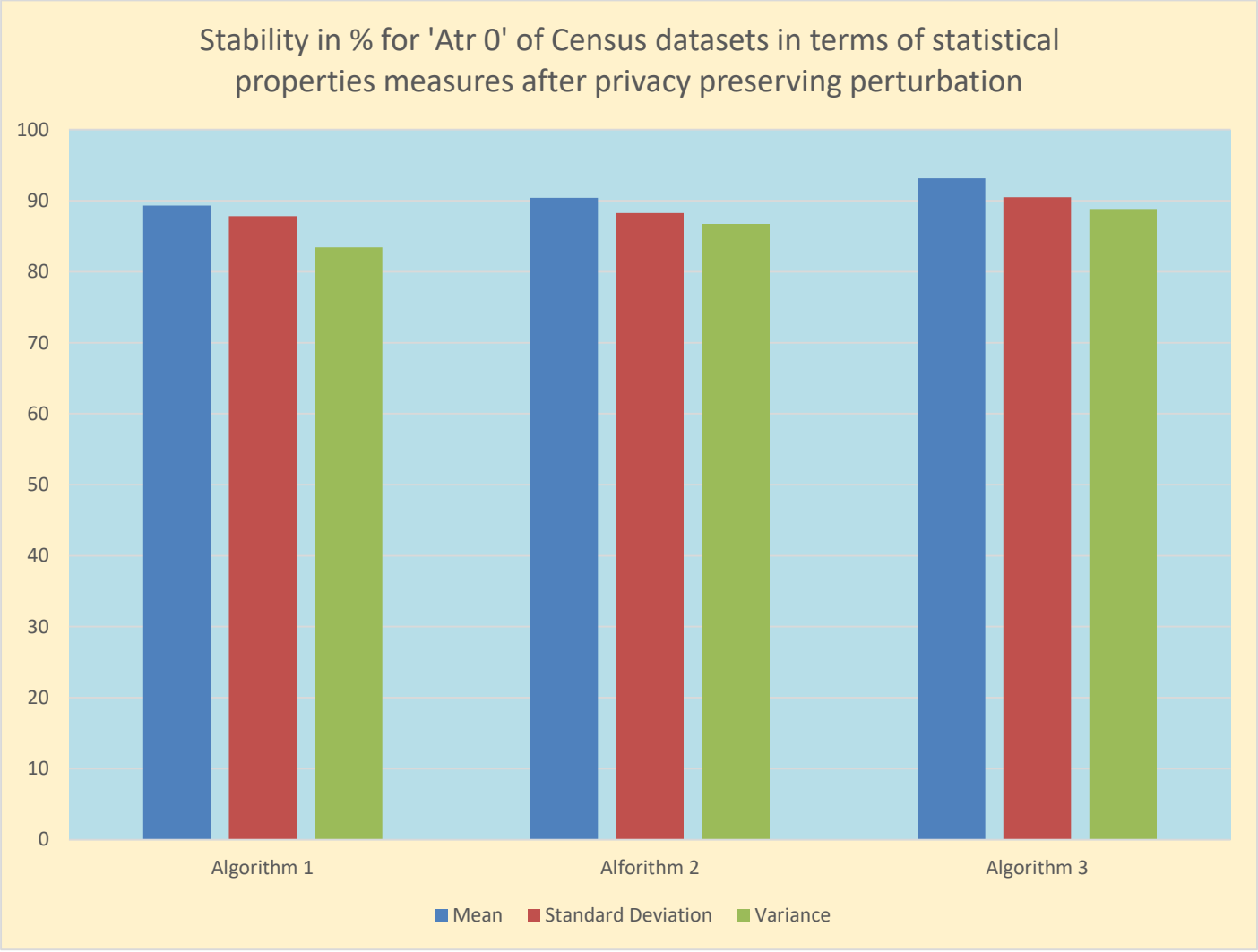


Figure 8: Stability in % for 'Atr 0' of Census datasets in terms of statistical properties measures after privacy preserving perturbation

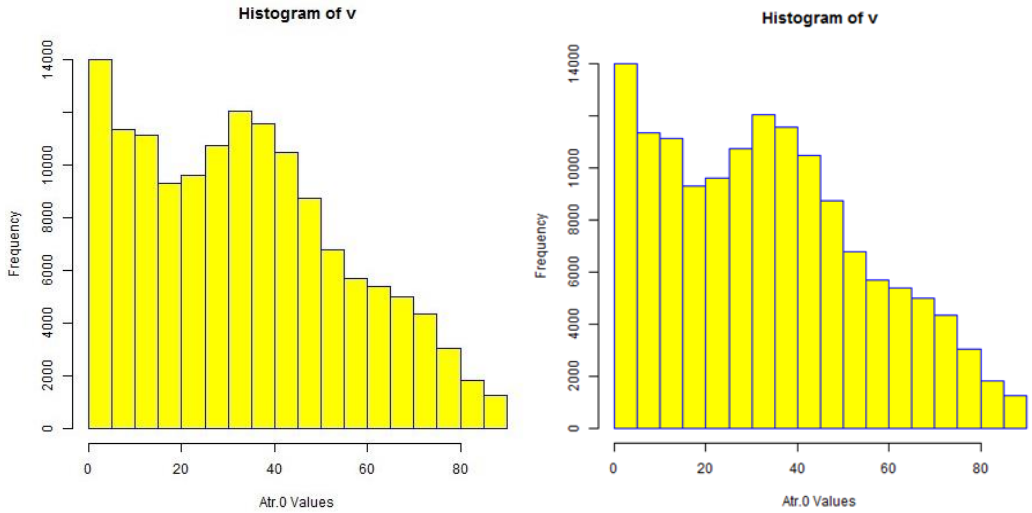


Figure 9: Histograms of 'Att 0' attribute of original and perturbed Census dataset using Algorithm 3

The Fig. 10, given below demonstrates the boxplot image of the trait 'Atr 0' of the original and changed probationary dataset Census. From the chart, it us identified that there are minor alterations in the statistical properties of the guesstimates of the original and adjusted datasets for the trait 'Atr 0' due to privacy conserving perturbation.

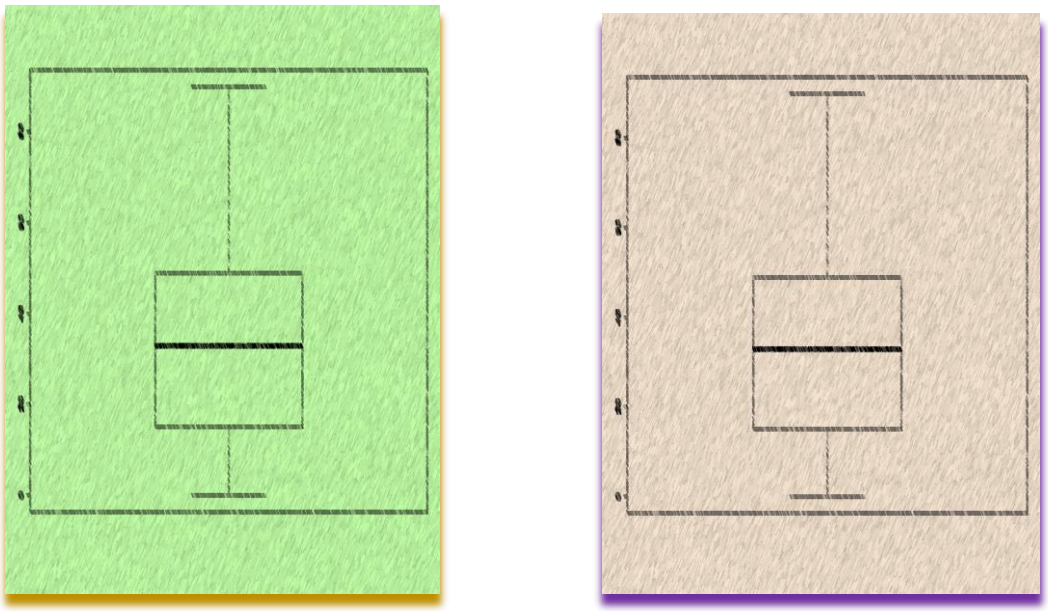


Figure 10: Boxplots of 'Att 0' attribute of original and perturbed Census dataset using Algorithm 3

The CFS feature selection algorithm was applied to choose attributes from every original and privacy-preserved dataset and the search procedure exploited in the trial was BestFirst. The CFS algorithm is filter-predicated and along these lines it does not team up with any classifier inside the selection strategy. Over fitting is abridged by a 10-fold cross-validation. BestFirst rehearses greedy hillclimbing to look for the region of attribute subsets and is upgraded with backtracking capabilities. BestFirst can seek in reverse once it starts with the complete set of traits or can pursuit forward once it starts with the vacant set of traits or inquiries toward every path at any point by demanding that all single trait augmentations and cancellations are contemplated at a specified point.

The number of traits chosen has been kept up in an ideal range, as the stability of the selection of features augments up to the ideal number of pertinent traits and after that declines. The statistic qualities, for example, mean, variance and standard deviation of both the original and the revamped

dataset numerical attributes are then estimated. The endeavours for statistical recitals are directed to the privacy conserved datasets for corroboration.

Feature selection stability guesstimates of the dataset Census and dataset Coil 2000 are gauged with the stability measure Kuncheva Index KI and the outcomes are recorded in Fig.11. In view of the KI, the greater cardinality esteem does not influence the stability of the feature selection and hence was exploited as a stability measure during the investigations. The stability of the selection of features together with the data utility is antagonistically linked to the discrepancy of the dataset, for instance, like the ruffling of the datasets.

The algorithms for privacy protection have created practically unswerving stable feature selection outcomes in light of the statistical properties of the numerical attributes of the bothered datasets. The Coil 2000 dataset has all the numerical traits, while the Census dataset has categorical and numerical traits. Furthermore, it was found from the repercussions that the Coil 2000 dataset is significantly more stable than the Census dataset because it just encompasses numerical traits.

The stability of the selection of features and the data utility are correlated. Since the feature selection stability repercussions for the privacy preservation algorithms are workable, the accuracy of the privacy conserved datasets remains nearly equivalent as before ruffling.

In the experimental studies, the accuracy results are given priority over other measures such as precision, recall, F1-score, ROC, AUC because there is tradeoff among privacy-preserving perturbation, feature selection stability, and accuracy. In this way, the planned algorithms for privacy protection have been seasoned with two utterly diverse datasets for privacy conservation, the stability of feature selection, and data utility. Moreover, the trial outcomes have demonstrated that the operation of the algorithms on investigational datasets prompts a stable selection of features with an unassailable accuracy. Table 2 condenses the measures of the steered experimentation on the trial datasets alongside the kinfolk with feature selection stability and accuracy.

Table 2: Summary of feature selection stability and accuracy measures for datasets Census and Coil 2000 for Algorithm 1, Algorithm 2 and Algorithm 3

Experimental Results	**Algorithm 1**		**Algorithm 2**		**Algorithm 3**	
	Census	**Coil 2000**	**Census**	**Coil 2000**	**Census**	**Coil 2000**
Feature selection stability exploiting Kuncheva Index KI	0.82	0.87	0.86	0.91	0.89	0.93
Overall accuracy before ruffling	73.78%	75.93%	73.78%	75.93%	73.78%	75.93%
Overall accuracy after ruffling	66.36%	68.94%	69.73%	71.62%	71.85%	72.61%
The Accuracy of chose features before ruffling	78.83%	81.79%	78.83%	81.79%	78.83%	81.79%

The Accuracy of chose features after ruffling	73.26%	75.92%	74.42%	77.86%	76.52%	78.72%

The measures of the investigational outcomes of the three algorithms are almost the same. Nevertheless, Algorithm 3 offers a slightly better recital than the other two algorithms, as displayed in the diagram below in Fig. 11. Conversely, the outcomes of Algorithm 2 are slightly better than Algorithm 1, because it rehearses the slicing technique that has better consequences than the data distortion techniques of Algorithm 1.

8. CONCLUSION

The present era's microdata are high-dimensional hence feature selection is considered as an indispensable dimension lessening approach. However, the stability of the selection of features is crucial for choppy datasets like privacy-conserved datasets as stability is generally data-centric. The privacy preserving data mining concerns both privacy and security of citizenry, as well as, the data utility. In order to protect the data that are sensitive to privacy, the dataset must be perturbed after exploiting a few methods that will disturb the stability of the feature selection, since the stability of the feature selection depends upon the data in greater extents. In particular, researches that keep up the stability of the feature selection for such choppy privacy preserved datasets must be explored as selection stability so as to correlate data utility in future. The present paper investigates the trade-off amongst ruffling of a dataset for the preservation of privacy, the stability of selection of features, and data utility for choppy privacy preserved datasets.

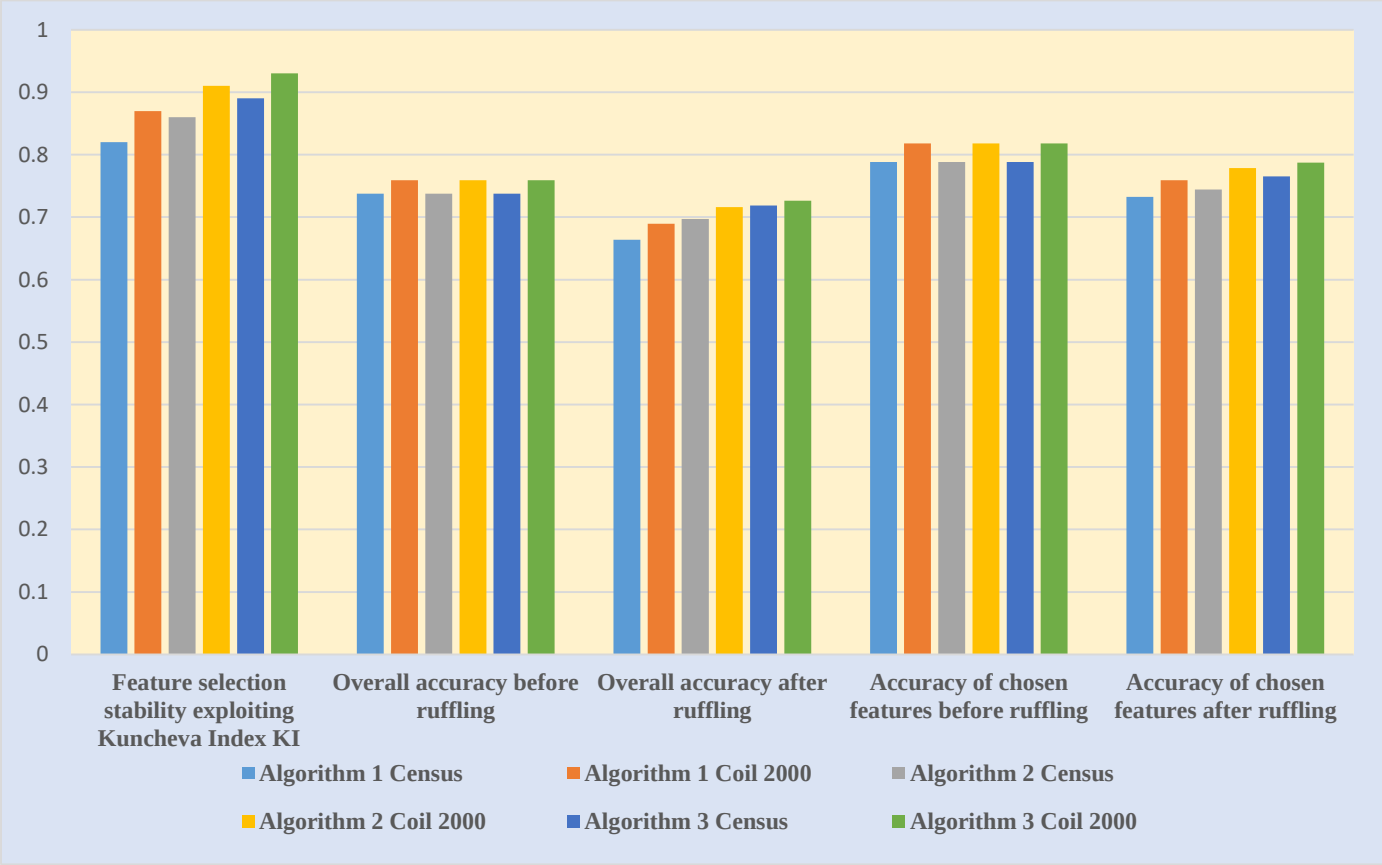


Figure 11: Feature selection stability and accuracy measures for datasets Census and Coil 2000 for Algorithm 1, Algorithm 2 and Algorithm 3

9. FUTURE WORK

In view of the findings of the present research paper it is concluded that the characteristics of dataset greatly influence the feature selection stability results, especially for choppy datasets like the privacy preserved datasets. Hence it is identified that the application of privacy-preserving techniques need not have much influence upon the characteristics of datasets. The future selection can be extended using artificial intelligence and the former can be applied for privacy-preserving techniques for better feature selection stability results and accuracy. Apart from accuracy, other measures such as precision, recall, F1-score, ROC, and AUC can also be analysed to evaluate the algorithms in diverse ways.

REFERENCES

1. Dr. Brijesh Kumar Bhardwaj, A Critically Review of Data Mining Segment: A New Perspective, International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No. 6, 2019 https://doi.org/10.30534/ijatcse/2019/50862019
2. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer, 2001.
3. H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Boston: Kluwer Academic Publishers, 1998.
4. Guyon and A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research, 3:1157–1182, 2003.
5. Chad A Davis, Fabian Gerick, Volker Hintermair, Caroline C Friedel, Katrin Fundel, Robert Kfner, and Ralf Zimmer, Reliable gene signatures for microarray classification: assessment of stability and performance, Bioinformatics, 22(19):2356–2363, Oct. 2006
6. Mark, A., Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer science, University of Waikato. http://www.cs.waikato.ac.nz/ mhall / thesis.pdf, 1998.
7. AlexandrosKalousis, Julien Prados, and Melanie Hilario, Stability of feature selectionalgorithms: a study on high-dimensional spaces, Knowledge and Information Systems, 12(1):95–116, http://link.springer.com/article/10.1007/s10115-006-0040-8, May 2007.
8. Zengyou He and Weichuan Yu, Stable feature selection for biomarker discovery, 2010.
9. Kalousis, A., Prados, J., and Hilario, M., Stability of feature selection algorithms, page 8, Nov. 2005
10. Lei Yu, Chris Ding, and Ste1ven Loscalzo, Stable feature selection via dense feature groups, In KDD ’08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 803– 811, New York, NY, USA, ACM, 2008
11. Salem Alelyani, Huan Liu., The Effect of the Characteristics of the Dataset on the Selection Stability, IEEE DOI 10.1109/International Conference on Tools with Artificial Xiniun Intelligence. 2011.167, 1082-3409/11, http:// ieeexplore.ieee.org/document/ 6103458, 2011.
12. Salem Alelyani, Zheng Zhao, Huan Liu., A Dilemma in Assessing Stability of Feature Selection Algorithms, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications. 2011.99, 978-0-7695- 4538-7/11, http:// ieeexplore.ieee.org/document/6063062, 2011.
13. Salem Alelyani, On feature selection stability: a data perspective, Doctoral Dissertation, Arizona State University, AZ, USA, ISBN: 978-1-303-02654-6, ACM Digital Library, 2013.
14. Barbara Pes, Feature Selection for High-Dimensional Data: The Issue of Stability, Proceedings of the 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2017), June 21–23, 2017.
15. Jundong Li, Huan Liu, Challenges of Feature Selection for Big Data Analytics, Special Issue on Big Data, IEEE Intelligent Systems, eprint arXiv:1611.01875, 2017
16. Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu, Feature Selection: A Data Perspective, ACM Comput. Surv. 50, 6, Article 94, 45 pages. DOI: https://doi.org/10.1145/3136625, Jan. 2018
17. Jianyu Miao, Lingfeng Niu, A Survey on Feature Selection,Elsevier B.V., doi: 10.1016/j.procs.2016.07.111, Procedia Computer Science 91, 2016.
18. Hall, M A., and Smith L A., Practical feature subset selection for machine learning, Proceedings of the 21st Australian Computer Science Conference, Springer.181-191, 1998.
19. Kevin Dunne, Padraig Cunningham, and Francisco Azuaje, Solutions to instability problems with sequential wrapper-based approaches to feature selection, Technical ReportTCD-CD-2002-28, Department of Computer Science, Trinity College, Dublin, Ireland, 2002.
20. L. I. Kuncheva, A stability index for feature selection, In Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi Conference: artificial intelligence and applications, pages 390 - 395, Anaheim, CA, USA, ACTA Press, 2007.
21. L. Burnett, K. Barlow-Stewart, A. Proos, and H. Aizenberg, The Gene Trustee: a universal identification system that ensures privacy and confidentiality for human genetic databases, Journal of Law and Medicine, 10(4):506-513, 2003.
22. Mar ValverdeLópez, Data Protection: A new Regulation in the European Union,International Journal of Advanced Trends in Computer Science and Engineering, Volume 5, No. 4, 2016 http://www.warse.org/IJATCSE/static/pdf/file/ijatcse03542016.pdf
23. Munir Ahmed, Lukman Sharif, Muhammad Kabir, Maha Al-Maimani, Human Errors in Information

Security, International Journal of Advanced Trends in Computer Science and Engineering, Volume 1, No. 3, 2012 http://warse.org/pdfs/ijatcse01132012.pdf

24. Xiniun, Q., MingkuiZong, An Overview of Privacy Preserving Data Mining, 1878- 0296, doi: 10.1016/Procedia Environmental Sciences 12, 2012.
25. Yousra Abdul Alsahib S. Aldeen, MazleenaSalleh and Mohammad Abdur Razzaque, A comprehensive review on privacy preserving data mining, SpringerPlus, DOI 10.1186/s40064-015-1481-x, 2015.
26. Ricardo Mendes, João P. Vilela, Privacy-Preserving Data Mining: Methods Metrics and Applications, IEEE Access, vol. 5, pp. 10562-10582, 2017.
27. Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy, Slicing: A New Approach for Privacy Preserving Data Publishing, Transactions on knowledge and data engineering, vol. 24, no. 3, pp.561-574, Mar. 2012.
28. Tiancheng Li, Ninghui Li, Injector: mining background knowledge for data anonymization, In IEEE 24th International Conference on Data Engineering (ICDE 2008), IEEE, Cancun, Mexico, 446–455, 2008.
29. P. MohanaChelvan and K. Perumal, Stable Feature Selection with Privacy Preserving Data Mining Algorithm, Springer International Conference on Advanced Informatics for Computing Research (ICAICR 2017), Springer Communications in Computer and Information Science (CCIS) Series, ISSN: 1865-0929, CCIS 712, pp. 227-237, DOI: 10.1007/978-981-10-5780-9_21.
30. Priti S. Patel, Dr. S.G. Desai, A Comparative Study on Data Mining Tools, International Journal of Advanced Trends in Computer Science and Engineering, Volume 4, No. 2, 2015 http://www.warse.org/ijatcse/static/pdf/file/ijatcse04422015.pdf
31. Alcalá-Fdez, A., Fernández, J., Luengo, J., Derrac, S., García, L. Sánchez, and Herrera, F., KEEL data-mining software tool: Dataset repository, integration of algorithms and experimental analysis framework, J. Multiple- Valued Logic Soft Comput., 17(2): 255–287, 2010.